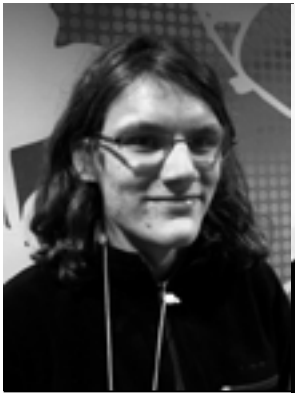
Viestinnän mittaaminen Big Datan avulla

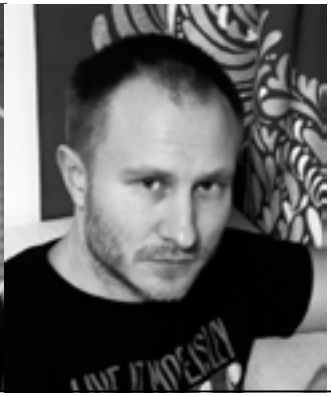
Nelimarkka, Matti

þÿ P r o C o m V i e s t i n n ä n a m m a t t i l a i s e t r y
2017

Nelimarkka , M & Sund , R T 2017 , Viestinnän mittaaminen Big Datan avulla . julkaisussa E Juholin & V Luoma-aho (toim) , Mitattava viestintä . , 7 , ProComma Academic , Vuosikerta.
þÿ 2 0 1 7 , P r o C o m V i e s t i n n ä n a m m a t t i l a i s e t r y , H e l s i n k i , S i v u t 7 4 - 8 5 .

http://hdl.handle.net/10138/311229
https://doi.org/10.31885/2018.00030





KETKÄ?
Matti Nelimarkka ja Reijo Sund

- *Matti Nelimarkka on Helsingin yliopiston jatko-opiskelija ja Aalto-yliopiston tutkija. Hänellä on sekä yhteiskuntatieteen että tietojenkäsittelytieteen tutkinnot. Hänen tutkimusaiheensa pureutuvat sosiaalisen laskennan alalle: hän kehittää ja tutkii erityisesti uusia vuorovaikutteisia sovelluksia sekä analysoi ihmisten välistä tietokonevälitteistä ja tietokonetuettua vuorovaikutusta. Nelimarkka on Rajapinta ry:n puheenjohtaja sekä perustajajäsen ja opettaa laskennallista yhteiskuntatiedettä Helsingin yliopistolla.*

Reijo Sund toimii Helsingin yliopiston yhteiskuntatieteiden menetelmäkeskuksen johtajana sekä professorina Itä-Suomen yliopistossa. Hän on soveltavan tilastotieteen dosentti, ja hänen tutkimuskiinnostuksensa ovat kohdistuneet ensisijaisesti isojen rekisteriaineistojen analysointiin liittyviin metodologisiin kysymyksiin sekä tilastolliseen tietojenkäsittelyyn niin yhteiskunta- kuin terveystieteisiin kuuluvissa sovelluksissa. Sund on osallistunut myös avoimeen lähdekoodiin perustuvien tilastollisten ohjelmistojen kehittämiseen.

VIESTINNÄN MITTAAMINEN BIG DATAN AVULLA

Viime aikoina monella yhteiskuntatieteiden alalla yhdeksi uudenlaiseksi ja lupaavaksi tutkimusmahdollisuudeksi on noussut massadatan (big data) käyttö, niin myös viestintätieteissä. Yleisellä tasolla uudenlainen tutkimustapa on kiinnostanut erityisesti sosiaalisen median palveluiden analyysissä, kuten Twitterin ja Facebookin kohdalla (esimerkiksi Zhang & Counts, 2015; Bruns & Highfield, 2013; Hemphill & Roback, 2014; Graham et al., 2013), mutta massadatan analyysiin kehitettyjä menetelmiä on käytetty myös perinteisen media-aineiston analyysiin (esimerkiksi Levy & Franklin, 2013; Burscher, Vliegenthart & De Vreese, 2015). Kohteena ovat olleet myös tekstipohjaiset aineistot – joihin rajaudumme tässä työssä – vaikkakin viime aikoina myös kuvien, videoiden sekä äänen analyysiin on kehitetty paljon uusia menetelmiä.

Riippuen puhujasta massadata saattaa viitata paitsi aineistoon myös menetelmällisiin lähestymistapoihin aineiston analyysissä. Mielestämme yhteiskuntatieteellisen tutkimuksen yhteydessä nämä uudet laskennalliset analyysimenetelmät ovat jopa kiinnostavampia kuin itse massadata. Analyysimenetelmiin keskittymällä vältämme varsin turhauttavan keskustelun massadatan koosta ja eroista ei-massadataan. Nämä menetelmät tukevat perinteisiä kyselytutkimuksen ja haastatteluaineistojen laadullisia ja määrällisiä analyysitapoja. Laskennallisen yhteiskuntatieteen piirissä suosittuja menetelmiä ovat olleet koneoppiminen, verkostoanalyysi sekä simulaatiomallit (Cioffi-Revilla, 2010). Lisäksi erityisesti datan visualisoinnin kautta vuorovaikutteisten järjestelmien käyttö tarjoaa analyyseihin mielekästä lisäsisältöä.

Tässä lyhyessä katsauksessa käsittelemme viestinnän mittaamista laskennallisilla menetelmillä. Mittaaminen keskittyy usein saavutetun viestinnän arvon mittaamiseen suhteessa kustannuksiin; mittareita voivat esimerkiksi olla uusien kommenttien lukumäärä sosiaalisessa mediassa (heijastelee aktiivisuutta) kuin uutisotsikoihin nousseet viestit (heijastelevat ulkoista näkyvyyttä). Aloitamme määrittelemällä, mitä tarkoitamme massadatalla ja kuvaamalla yleisellä tasolla sen analysointimenetelmiä. Tämän jälkeen käsittelemme neljä erilaista näkökulmaa mittaamiseen: määrällisen, laadullisen, tilastollisen sekä ennustavan. Päätämme katsauksen arvioimalla, mihin massadata ja sen menetelmät soveltuvat sekä eivät sovellu viestinnän mittaamisen analyy-

sissä ja kuinka massadataa voi hyödyntää viestintäalan ammattilaisena.

Big data ja analyysimenetelmiä

Viimeisten muutaman vuoden aikana termi ”Big data” on noussut yhdeksi muotitermiksi, jota ryöstöviljellään miltei alalla kuin alalla (Sund, 2015). Usein tuntuu lisäksi olevan epäselvää, mitä termillä missäkin yhteydessä tarkoitetaan. Suomenkielisen Wikipedian mukaan ”Big data on erittäin suurten, järjestelemättömien, jatkuvasti lisääntyvien tietomassojen keräämistä, säilyttämistä, jakamista, etsimistä, analysointia sekä esittämistä tilastotiedettä ja tietotekniikkaa hyödyntäen.” (Wikipedia, 2017.) Kyseisessä määritelmässä on pyritty tietynasteiseen arvovapauteen, mutta muotitermiksi nouseminen on vaatinut uskoa massadatan huomattavaan taloudellisen hyödyntämisen potentiaaliin. Massadatasta puhuttaessa törmätäänkin usein hehkutukseen, jossa erilaiset (yritysten) tietovarannot luvataan taianomaisesti muuttaa (liiketoimintaa tukevaksi) hyödylliseksi informaatioksi. Ennen massadataa (2010-luku) käytännössä samasta asiasta puhuttiin ainakin liiketoimintatiedon hallinnan (*business intelligence,* 2000-luku), tiedonlouhinnan (*data mining,* 1990-luku) ja asiantuntijajärjestelmien (*expert systems,* 1980-luku) nimillä, mutta ne eivät onnistuneet täyttämään monia suurista lupauksistaan. Keskeinen syy siihen, etteivät ne nousseet valtavirran analyysiksi, oli liiallinen usko datan objektiivisuuteen sekä keskittyminen teknologiaan eikä varsinaisiin empiirisen tutkimuksen perustavanlaatuisin haasteisiin.

On tarpeen korostaa, ettei datan määrää lisäämällä voida ohittaa empiirisen tutkimuksen perusoletuksia tai rajoituksia. Erilaiset käsitykset aineistojen tavasta heijastella todellisuutta johtavat myös erilaisiin metodologisiin valintoihin. Jos mitattavat asiat ovat suoraan havaittavia, kuten esimerkiksi lähetettyjen viestien lukumäärä, on niitä varsin helppoa tarkastella siten, että jokainen ymmärtää asian samalla tavalla. Jos taas ollaan kiinnostuneita toisentyyppisistä kysymyksistä, metodologinen tilanne on monimutkaisempi. Esimerkiksi kun kysytään, miksi joku viesti on lähetetty, sitä harvoin pystytään mittaamaan käytettävissä olevasta aineistosta millään itsestään selvällä tavalla. Tässä suhteessa yhteiskuntatieteet ovat tyypillisiltä tutkimuskysymyksiltään eriytyneet suorempaan havainnointiin perustuvista tieteistä. Tämä on ongelmallista siinä mielessä, että menetelmien soveltaminen tilanteissa, joissa tutkimuskysymys tai aineisto ei ole sopusoinnussa menetelmän kannalta oleellisten metodologisten oletusten kanssa, johtaa usein kestämättömiin tulkintoihin ja pahimmillaan karkeisiin virhepäätelmiin. Lisähaasteita aiheuttaa vielä se, että käytännössä kaikki massadatat ovat luonteeltaan sekundaarisia, eli niitä ei alun perin ole tuotettu nimenomaan kyseessä olevaa tutkimusta varten. Varsinkin sosiaalisen median aineistojen osalta on myös ongelmallista, että yhteiskunnalliset ilmiöt ja itse sosiaalinen media voivat vaikuttaa hyvin nopeasti syntyvän aineiston luonteeseen, jolloin tutkimuksen tai vastaavan asetelman toistaminen myöhemmin ei välttä-

mättä enää onnistu - esimerkiksi Twitterin keskusteluissa yksilöiden Twitterin käyttö sekä tulkinnat aiheiden kiinnostavuudesta muokkaavat tämän median sisältöä (Jungherr, Schoen & Jürgens, 2016).

Käytännössä nämä rajoitukset tarkoittavat sitä, että kysymyksenasettelua joudutaan muokkaamaan sellaiseksi, että siihen pystytään vastaamaan käytettävissä olevalla aineistolla. Tämä on vaikeaa, kun tarkastellaan moniulotteisia ilmiöitä kuten viestinnän mittaamista, ja vaatii varsin laajaa taustatietämystä aiheesta ennen kuin edes yritetään soveltaa laskennallista analyysiä (Boyd & Crawford, 2012). Ilman ilmiötä ja tarkasteltavan aineiston syntyhistoriaa koskevaa taustatietämystä päädytään täysin aineistolähtöiseen tutkimukseen, joka ei yleensä milloinkaan ole tarkoituksenmukaista. Taustatietämyksen lisäksi tarvitaan osaamista myös laskennallisten menetelmien käytöstä ja isojen aineistojen käsittelystä, joten onnistuneeseen tutkimukseen tarvitaankin välttämättä yhteiskuntatieteen ja laskennallisen osaamisen yhdistämistä (Halavais, 2015).

Laskennallinen näkökulma viestinnän mittaamiseen

Yksinkertaisin lähestymistapa viestinnän mittaamiseen on laskea lukumääriä. Laskennalliset menetelmät ovat erinomainen tapa tuottaa lukumäärätietoa laajoista aineistoista, erityisesti jos aineistolle on tarpeen tehdä jotain esikäsittelyä – kuten muuttaa sanat perusmuotoon – ennen laskentaa. Yleisesti erilaiset kuvailevat tunnusluvut ovat yksinkertaisin, usein tosin alkeellinen, tapa mitata viestintää isoista aineistomassoista. Esimerkiksi viestien määrä, jakauma ajan suhteen sekä viestien tyypit (kommentteja, vastauksia, linkkien jakoja yms.) ovat yleisesti käytössä olevia mittareita (esimerkiksi Marttila et al., 2016; Merry, 2014; Graham et al., 2013). Usein tulokset ovat kiinnostavia, mutta varsin vaillinaisia – ilmiötä pystytään kyllä kuvailemaan, mutta ei välttämättä täsmällisemmin erottelemaan ilmiön laatua.

Myös verkostoanalyysia käytetään usein viestinnällisen datan analyysiin. Verkostoanalyysissä ilmiö kuvataan toimijoiden välisiksi suhteiksi, verkoksi jossa havainnoidaan toimijoiden (‘solmujen’) välille muodostuvia yhteyksiä (‘kaaria’). Solmujen ja kaarien kautta voidaan laskea useita verkkoa kuvaavia tunnuslukuja, kuten aste, joka mittaa solmuun tulevien kaarien määrää, tai havainnoida millaisia yhteisöjä verkossa syntyy. Perinteinen verkostoanalyysin sovellus viestinnän mittaamisessa on tutkia sosiaalisen median vuorovaikutussuhteita tietyn tapahtuman aikana. Esimerkiksi Larsson (2013) on tutkinut, kuinka katsojat ja esiintyjät viestivät toisilleen television puheohjelmien aikana. Samoin Bruns ja Highfield (2013) tarkastelevat verkostoanalyysin menetelmiä käyttäen vuorovaikutusta vaalien alla eri toimijaryhmien (ehdokkaat, media, kansalaiset jne.) välillä. Tämänkaltaisella lähestymistavalla voidaan arvioida, onnistuuko viestintä luomaan monensuuntaista vuorovaikutusta vai tapahtuuko vuorovaikutus vain viestijän ja yleisön välillä, jolloin vuorovaikutuksellisuudesta huolimatta ei välttämättä synny laajaa keskustelua.

Sisällön analyysi isojen aineistojen aikana

Edellä kuvattua määrällistä tapaa on kuitenkin helppo kritisoida viestin varsinaisen sisällön huomiotta jättämisestä. Vaikka sisällön ja merkitysten analyysi on selvästi haastavampaa kuin lukumäärien tarkastelu, myös koneelliseen sisällön analyysiin on kehitetty menetelmiä. Yksi lähestymistapa on sentimenttianalyysi (esim. Thelwall et al., 2010), jonka avulla voidaan arvioida viestin sisältöä positiivisella ja negatiivisella akselilla. Laskennalliset menetelmät ovat saavuttaneet paljon suosiota, sillä niitä käyttämällä saadaan näennäisesti analysoitua sellaisiakin isoja aineistoja, joiden sisältämän tekstimäärän lukeminen veisi vuosikausia. Monissa tilanteissa näiden menetelmien toiminta voi kuitenkin olla ristiriitaista – esimerkiksi ironian tulkinta on haastavaa niin ihmisille kuin tietokoneille – tai viestin negatiivinen ilmaus koskeekin jotain muuta kuin viestissä mainittuja viestijöitä. Lisäksi sentimenttianalyysi perustuu usein sanalistoihin positiivisista ja negatiivisista ilmaisuista; kuitenkin näiden sanalistojen validiteetti – eli kuinka hyvin sanalistat kuvaavat positiivisia ja negatiivisia tunteita, erityisesti englannista suomeksi käännetyissä listoissa – voi olla todella huono. Laskennallisen sisällönanalyysin menetelmiä käytettäessä tulisikin aina edes jossain määrin pyrkiä arvioimaan koneellisen tulkinnan tarkoituksenmukaisuutta lukemalla joitain alkuperäisiä tekstejä ja vertaamalla koneen tulkintaa inhimilliseen tulkintaan (Grimmer & Stewart, 2013). Muuten käy helposti niin, että aletaan tehdä tulkintoja vain laskennallisen menetelmän tuloksista eikä varsinaisesta sisällöstä. Silloin päädytään usein sellaiseen virhepäätelmään, jossa saadaan kyllä laskennallisessa mielessä oikeita vastauksia, mutta ei niihin kysymyksiin joihin haluttaisiin.

Jonkin teoreettisesti pohjustetun kehikon käyttäminen viestien luokitteluun on harvemmin viestinnän määrällisessä mittauksessa sovellettu mutta mielekäs apuväline viestien lukumäärän laskemisen yhteyteen. Yleisesti tällaista lähestymistapaa kutsutaan määrällisissä analyyseissa datan rikastamiseksi, eli käytännössä hyödynnetään analyyseissa taustatietämystä, jota ei aineistosta sinänsä suoraan löydy mutta jonka mukaisesti aineiston voi kuitenkin epäsuorasti olettaa käyttäytyvän. Tätä voidaan tehdä monella tavalla, mutta yksi todella kätevä laskennallinen keino on käyttää ohjattua koneoppimista: ensin opetetaan tietokonetta esimerkkien avulla, minkä jälkeen tietokone voi käyttää oppimaansa luokittelua hyvin nopeasti suuriinkin aineistoihin. Opettamalla tietokonetta luokittelemaan sisältöjä olemassa olevien (mahdollisesti teoreettisesti mielekkäiden) luokitusten kautta voidaan tehostaa aineiston analyysiä merkittävästi. Esimerkiksi Stromer-Galley et al. (2016) toteuttivat Yhdysvaltojen 2016 presidentinvaaleihin sovelluksen, joka luokitteli ehdokkaiden sosiaalisen median viestejä niiden viestintätarkoituksen perusteella, esimerkiksi tunnisti hyökkäävyyttä sekä imagon ylläpitoa.

Monissa tapauksissa on mahdollista käyttää analyysiin myös etäisyysmittoihin

perustuvaa ryhmittelyanalyysia tai ohjaamatonta koneoppimista, joissa pyritään puhtaasti laskennallisin keinoin ryhmittelemään aineistoa. Tällöin ei siis ole tarpeen opettaa koneelle luokittelua (kuten ohjatussa koneoppimisessa), vaan kone muodostaa itsenäisesti mahdollisia luokitteluja ja ihmisen tehtäväksi jää näiden tulkinta. Viime aikoina aihemallinnus (*topic modeling*) on kerännyt suosiota (Blei, 2012). Aihemallinnuksessa tietokone ryhmittelee aineiston sanoja ryhmiin. Jokaiselle "dokumentille" (=yhdelle tekstiaineistoille) lasketaan todennäköisyys kuulua tiettyyn aiheeseen eli ryhmittelyn mukaiseen termi- ja dokumenttijoukkoon. Aiheet ovat koneen havaitsemia yleisiä sanoja, jotka esiintyvät useimmiten joissain tietyissä dokumenteissa. Aiheille annetaan myöhemmin merkitys tarkastelemalla näitä sanoja. Menetelmää on sovellettu Suomessa esimerkiksi vaalijulkisuuden havainnointiin. Vuoden 2015 vaalien lehdistöjulkisuudessa voitiin koneellisesti löytää 40 erilaista politiikkaan liittyvää aihekokonaisuutta (esimerkiksi energiapolitiikka, lapsiperheet sekä valtiontalous). Nämä aihekokonaisuudet löydettiin aineistosta ilman etukäteen annettua luokittelua (Nelimarkka et al., arvioitavana).

Tilastolliset menetelmät ja big data

Yllä kuvatut menetelmät ovat kuitenkin luonteeltaan kuvailevia, ne eivät sellaisenaan kerro vielä syy-seuraussuhteista – jotka ovat usein mielenkiinnon kohteena myös viestinnän mittauksessa. Syy-seuraussuhteita tarkasteltaessa käytetään tyypillisesti konfirmatorista lähestymistapaa, jossa perinteiseen määrällisen tutkimuksen tapaan asetetaan ensin hypoteesi ja tarkastellaan sitten sen toteutumista tilastollisilla testeillä. Esimerkiksi Laaksonen et al. (2017) uskoivat, että ehdokkaat ovat hyökkäävämpiä, jos he kommentoivat kilpailevan puolueen ehdokkaan sisältöjä Facebookissa kuin jos he kommentoivat oman puolueensa ehdokkaan sisältöjä. Negatiivinen viestintä tunnistettiin laskennallisesti sentimenttianalyysillä, ja viestien sentimenttien jakaumaa tarkasteltiin perinteisellä tilastollisella testaamisella.

Syy-seuraussuhteen tarkastelemiseksi ilmiö tyypillisesti operationalisoidaan aiemman teorian ja tutkimuskirjallisuuden avulla. Operationalisoinnissa yhdistetään tutkimuksellinen käsite erilaisiin mittareihin, esimerkiksi kyselytutkimuksessa joukko kysymyksiä operationalisoidaan tarkoittamaan yhteiskuntatieteellisesti kiinnostavaa käsitettä. Tämä operationalisointi ja teoreettinen pohdinta voidaan myös laskennallisten menetelmien hengessä formuloida algoritmiksi, tietokoneohjelmaksi, joka tekee aineistolle mielekkään operationalisoinnin. Esimerkiksi Digivaalit 2015 -hankkeessa ehdokkaan vaikuttavuutta perinteisen median agendaan mitattiin tarkastelemalla, mainitsiko ehdokas sosiaalisessa mediassa poliittisesti mielenkiintoisen ilmiön ennen kuin siitä puhuttiin perinteisessä mediassa. Kyseessä on siis aikasarja-analyysin eräs sovellus, jolla mitattiin ehdokkaiden mahdollisuuksia tai potentiaalia vaikuttaa median agendaan. Näin laskettiin vaikuttajaindeksiä ehdokkaille. Myöhemmin hankkeessa tarkasteltiin taustatekijöiden, kuten iän,

sukupuolen sekä puolueaseman vaikutusta vaikuttajaindeksiin regressioanalyysissä ja havaittiin esimerkiksi eduskunnassa olevilla puolueilla olevan suuremmat vaikuttaja-indeksit median agendaan (Nelimarkka et al., arvioitavana).

Ennustaminen ja ennakointi massadatalla

Massadatalla pyritään usein ennustamaan tulevaisuutta. Klassinen esimerkki on Googlen aikanaan kehittämä flunssan ennustaminen ihmisten hakutermejä tarkastelemalla. Vaikka myöhemmin kyseinen tutkimustulos havaittiin pätemättömäksi (tiivistelmä kritiikistä, esim. Lazer et al., 2014), se on edelleen usein esillä. Esimerkiksi Suomessa samankaltaista lähestymistapaa on käytetty työttömyysasteen ennustamiseen. Ideana näissä ennustavissa malleissa on käyttää kerättyä historiallista massadataa ja yhdistää näitä aineistoja havaittuihin käytösmalleihin.

Ennustaminen on usealle viestinnän alalle mielekäs alue. Esimerkiksi markkinoinnissa tulevaisuuden ennustaminen voisi usein olla hyödyllistä kuluttajien tarpeiden havaitsemiseksi. Vaalien tuloksen ennustamiseen voidaan käyttää vaikkapa sosiaalisen median viestintää, millä saavutetaan usein melko tarkkoja tuloksia. Tosin, kuten Gayo-Avello (2013) huomauttaa, ei näissä tutkimuksissa – lupauksista huolimatta – ennusteta vaalitulosta, vaan usein enemmänkin kehitetään malleja, jotka selittävät edellisten vaalien tuloksia. Eli esimerkiksi viestien määrään, niiden sentimentteihin ja viestimäärän muutoksiin perustuva selittävä malli estimoidaan käyttäen edellisiä vaaleja kuvaavaa dataa ja uskotaan, että samanlainen malli voisi päivitetyllä datalla toimia myös seuraavien vaalien tarkastelussa.

Mihin laskennalliset menetelmät soveltuvat?

Periaatteessa laskennalliset menetelmät ovat hyvin samankaltaisia kuin perinteisemmätkin tilastotieteen menetelmät. Massadatan kanssa käytettäväksi kehitettyinä niillä on kuitenkin muutamia keskeisiä etuja. Ne skaalautuvat ja automatisoituvat, eli aineiston määrällä ei ole erityistä merkitystä käsittelyn kannalta ja samankaltainen analyysiprosessi voidaan suorittaa yksinkertaisesti uudelleen. Toisaalta laskennallisilla menetelmillä ja massadatalla on oikein käytettynä mahdollista tarkastella asioita toisenlaisesta näkökulmasta, joka voi tarkentaa tunnettua kuvaa ilmiöistä tai jopa paljastaa täysin uudenlaisia ilmiöitä. Tämän vuoksi niitä onkin luonnehdittu yhteiskuntatieteiden mikroskoopiksi (Lazer et al., 2009).

Skaalautuminen ja automatisoituminen mahdollistavat myös viestinnän reaaliaikaisen analyysin. Yllä mainitsemamme Yhdysvaltojen presidentinvaalianalyysi (Stromer-Galley et al., 2016) esimerkiksi tarjosi toimittajille mahdollisuuden tarkastella ehdokkaiden Twitter-aktiivisuutta päivittäin ja käyttää tätä omissa uutisissaan. Samoin erilaiset vuorovaikutteiset järjestelmät voivat hyödyntää koneellisesti tehtyä luokittelua. Kaye et al. (2012) halusivat parantaa yrityksen työntekijöiden tietoisuutta tuotteista käytävästä keskustelusta sosiaalisessa mediassa. He rakensivat

kahvihuoneeseen julkisen näytön, missä näytettiin tuotteita käsitteleviä positiivisia sekä negatiivisia viestejä sekä positiivisten ja negatiivisten viestien määrää. Näin jokainen kahvihuoneen käyttäjä tuli tietoiseksi yrityksen senhetkisestä maineesta ja "pöhinästä" sosiaalisessa mediassa.

Suuret lupaukset uusista analyysimahdollisuuksista ja löydöksistä massadatan ja laskennallisten menetelmien avulla ovat kuitenkin toistaiseksi jääneet varsin epämääräisiksi kokeiluiksi. Tällä viitataan erityisesti ohjaamattomaan koneoppimiseen ja sen sovelluksiin yhteiskuntatieteissä, jolloin siis suurten lupausten toteutuminen olisi eräänlainen aineistovetoisen (vs. teoriavetoisen) tutkimuksen paluu. Entuudestaan tuttua laajaa keskustelua on herättänyt kysymys teorian merkityksestä ja tarpeesta ilmiöiden tulkinnassa, niin akateemisessa kuin käytännöllisemmässä ammattitoiminnassa (kansainvälisestä keskustelusta, esimerkiksi Halavais, 2015; Kitchin, 2014; Boyd & Crawford, 2012). Tätäkin oleellisempi keskustelu liittyy kuitenkin epistemologisiin ja metodologisiin periaatteisiin; keskustelun seurauksena määrällinen yhteiskuntatieteellinen tutkimus aikaisemmin leimattiin positivistiseksi ja pinnalliseksi. Tämä suuntaus johti aikoinaan laadullisten lähestymistapojen korostumiseen, mutta se joutuu nyt käytännössä mukautumaan tilanteeseen, jossa pitäisi pystyä analysoimaan myös sekundaarisia massatekstiaineistoja. Emme usko emmekä toivo näiden uusien laskennallisten menetelmien korvaavan välttämätöntä yhteiskuntatieteelliseen tutkimukseen kuuluvaa teoriakeskustelua, mutta toivomme niiden johtavan tarkoituksenmukaisella tavalla määrällisten menetelmien arvostuksen palautumiseen ja niiden oikeutetun käytön yleistymiseen.

Yksi erittäin potentiaalinen mahdollisuus on pyrkiä tukemaan massadatalla sekä laskennallisilla menetelmillä muilla tavoin saatuja havaintoja. Esimerkiksi etnografiaa voidaan käyttää yhdessä laskennallisten menetelmien kanssa tukemaan aineiston keräämisessä ja analyysissä sekä ohjaamaan havaintojen muodostamisessa (Laaksonen et al., 2017). Myös sekundaaristen aineistojen ymmärtämiseen tähtäävälle dataetnografialle olisi tarvetta.

Viime aikoina on tapahtunut merkittävää kehitystä massadatan soveltamisessa myös muihin kuin tekstiaineistoihin, esimerkiksi kuviin. Kuvista voidaan yllättävän tarkasti tunnistaa sen sisältöjä (objekteja) ja luokitella kuvia näiden sisältöjen suhteen. Samoin esimerkiksi videoanalyysi perustuu tämänkaltaisen kuva-analyysin soveltamiseen, mutta sarjalle kuvia. Näiltäkään osin massadatan ja laskennallisten menetelmien mahdollisuuksia ei tule siis sivuuttaa.

Mitä massadatalla ei voi mitata?

Eräs massadataan liittyvä haave on korvata perinteiset kyselytutkimukset nopeammilla ja kustannustehokkaammilla, esimerkiksi verkkokeskusteluiden aineistoihin perustuvilla lähestymistavoilla (Gonzalez-Bailon & Paltoglou, 2015; Jungherr et al., 2016). Yhteiskuntatieteissä ollaan kuitenkin oltu yleisesti varsin skeptisiä tämän suhteen, koska verkkokeskustelut eivät ole oikein missään mielessä edustava otos; ne

perustuvat henkilön päätökseen ilmaista mielipiteensä sosiaalisessa mediassa (Jungherr, Schoen & Jürgens, 2016). Samoin esimerkiksi tapahtumien osallistujien määrän mittaaminen sosiaalisen median aktiivisuuden perusteella kärsii vastaavasta ongelmasta; sosiaalisen median aineisto voi mitata vain niitä henkilöitä, jotka haluavat kertoa olevansa tapahtumassa sosiaalisessa mediassa. Samankaltaisia ongelmia voi esiintyä myös muissa aineistoissa, esimerkiksi yrityksen aineisto kuluttajan ostoksista ei sisällä kilpailevan kauppaketjun aineistoa. Eräs keino onkin välttää tulosten yleistämistä tietyn ympäristön ulkopuolelle; esimerkiksi analyysiä Twitteristä ei ole syytä yleistää Twitterin ulkopuolelle.

Yllä käsittelimme jo kahta muuta haastetta, aineistojen sekundaarisuutta – eli sitä, että ne on tyypillisesti muodostettu muihin kuin ennalta tunnettuihin tutkimustarkoituksiin – kuin myös sitä, että monet yhteiskunnalliset ilmiöt muuttavat nopeasti muotoaan tavalla, joka heijastuu myös aineistoihin. Nämä rajoitukset eivät välttämättä haittaa kontekstiin tiukasti sidotuissa kuvailevissa tutkimuksissa, mutta jos niitä ei oteta tai pystytä ottamaan riittävällä tavalla huomioon tutkimuksessa, on vaarana ajautua joksikin aikaa tilanteeseen, jossa joka tuutista alkaa tulla huonolaatuisia, puhtaasti menetelmien ja aineistojen ehdoilla tuotettuja selvityksiä kunnollisen oivaltavan ja uutta luovan yhteiskuntatutkimuksen sijaan.

Big data ja yhteisöviestintä – mitä seuraavaksi?

Ensimmäinen – ja mielestämme suurin – haaste massadatan soveltamisessa on pystyä muotoilemaan kysymys, joka tukee viestintätarpeita ja johon voi uskottavasti vastata massadatan avulla. Useilla yrityksillä on tarjolla esimerkiksi sosiaalisen median analytiikkaan välineitä, mutta niiden kohdalla on syytä olla tarkkaavainen: kuvaavatko mittarit sitä mitä niiden pitäisi kuvata, ja onko lähestymistapa uskottava. Esimerkiksi mainitsemamme sentimenttianalyysivälineet ovat usein, erityisesti suomeksi, vielä alkutekijöissään.

Valmiiden välineiden etu on niiden kustannustehokkuus; toinen vaihtoehto on pyrkiä saamaan yrityksen sisälle viestinnällisen massadatan osaamista. Osaamista voi ostaa useilta ohjelmistoyrityksiltä, mutta myös ilmaista koulutusta laskennallisen datan analyysiin on tarjolla – esimerkiksi MOOC-alustoilla verkkokursseina. Jos tätä taitoa pyritään kehittämään yrityksessä, kannattaa turvautua ilmaisiin ja avoimiin data-analyysiympäristöihin ja datan analyysityökaluihin, kuten R-ohjelmointikieleen tai Python-kieleen ja sen scikit-learn-paketteihin. Molempiin näistä on saatavissa apua ja tutoriaaleja useista internetlähteistä.

Lopuksi

Kuten esimerkkimme näyttivät, laskennalliset menetelmät ovat monipuolinen ja uusia mahdollisuuksia avaava lähestymistapa viestinnän mittaamiseen. Tekstiaineistojen lisäksi löytyy vaihtoehtoisia tapoja niin kuvien, videoiden kuin tilastojenkin analyy-

siin. Suurin haaste liittyykin alati kasvaviin vaatimuksiin erilaisista tutkimustaidoista, joita onnistuneeseen analyysiin tarvitaan. Kuten Kitchin (2014) huomautti, myös yhteiskuntatieteellistä analyysia tarvitaan laskennallisen analyysin tekemiseksi yhteiskuntatieteissä. Tämä vaatii sekä tiivistä yhteistyötä menetelmäosaajien ja substanssiosaajien välillä että vähintäänkin yhteisen keskustelun mahdollistavaa peruskäsitystä niin menetelmistä kuin substanssistakin (Sund, 2003).

Kiitokset

Matti Nelimarkka kiittää Koneen Säätiötä sekä TEKESiä (hanke: Smarter Social Media Analytics) rahoituksesta tutkimukselle.

Kiitämme myös kahta nimetöntä arvioitsijaa artikkelin kommenteista, jotka selkeyttivät ja kehittivät artikkelia.

Näistä voit aloittaa

Grimmer, J. & Brandon M. S. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*(3): 267–297.

Jungherr, A., Harald S. & Pascal J. (2016). The Mediation of Politics through Twitter: An Analysis of Messages Posted during the Campaign for the German Federal Election 2013. *Journal of Computer-Mediated Communication 21*(1): 50–68.

KIRJALLISUUS

Blei, David M. (2012). Probabilistic Topic Models. *Communications of the ACM* 55(4): 77-84.

Boyd, Danah & Kate Crawford (2012). Critical Questions for Big Data. *Information, Communication & Society* 15(5). Routledge: 662–79.

Bruns, Axel & Tim Highfield (2013). Political Networks on Twitter. *Information, Communication & Society* 16(5). Routledge: 667–691.

Burscher, B., Vliegenthart, R. & De Vreese, C. H. (2015). Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? *The ANNALS of the American Academy of Political and Social Science* 659(1): 122–31.

Cioffi-Revilla, C. (2010). Computational Social Science. *Wiley Interdisciplinary Reviews: Computational Statistics* 2(3): 259–71.

Gayo-Avello, D. 2013. A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. *Social Science Computer Review* 31(6): 649–79.

Gonzalez-Bailon, S. & Paltoglou, G. (2015). Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. *The ANNALS of the American Academy of Political and Social Science* 659(1): 95–107.

Graham, T., Broersma, M., Hazelhoff, K. & van 't Haar, G. (2013). Between Broadcasting Political Messages and Interacting With Voters. The Use of Twitter during the 2010 UK General Election Campaign. *Information, Communication & Society* 16(5): 692–716.

Grimmer, J. & Brandon M. S. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21(3): 267–97.

Halavais, A. (2015). Bigger Sociological Imaginations: Framing Big Social Data Theory and Methods. *Information, Communication & Society* 4462 (June). 1–12.

Hemphill, L. & Andrew J. R. (2014). Tweet Acts: How Constituents Lobby Congress via Twitter. Teoksessa *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '14,* 1200–1210. New York, New York, USA: ACM Press.

Jungherr, A., Schoen, H. & Jürgens, P. (2016). The Mediation of Politics through Twitter: An Analysis of Messages Posted during the Campaign for the German Federal Election 2013. *Journal of Computer-Mediated Communication* 21(1): 50–68.

Jungherr, A., Schoen, H., Posegga, O. & Jürgens, P. (2016). Digital Trace Data in the Study of Public Opinion. An Indicator of Attention Toward Politics Rather Than Political Support. *Social Science Computer Review,* (online-first).

Kaye, J., Lillie, A., Jagdish, D., Walkup, J. Parada, R. & Mori, K. (2012). Nokia Internet Pulse. *Proceedings of the 2012 ACM Annual Conference Extended Abstracts on Human Factors in Computing Systems Extended Abstracts - CHI EA* '12: 829-844.

Kitchin, R. (2014). Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society* 1(1): 1–12.

Laaksonen, S.-M., Nelimarkka, M., Tuokko, M., Marttila, M Kekkonen, A., Villi, M. (2017). Working the fields of big data: Using big-data-augmented online ethnography to study candidate-candidate interaction at election time. *Information Technology & Politics* 14(1):1-22.

Larsson, A. O. (2013). Tweeting the Viewer - Use of Twitter in a Talk Show Context. *Journal of Broadcasting & Electronic Media* 57(2): 135–52.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science 343* (6167): 1203–1205.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N. et al. (2009). Social Science. Computational Social Science. *Science (New York, N.Y.)* 323: 721–23.

Levy, K. E. C. & Franklin, M. (2013). Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry. *Social Science Computer Review 32* (December): 182–94.

Marttila, M., Laaksonen, S-M., Kekkonen, A., Tuokko, M. & Nelimarkka, M. (2016). Digitaalinen Vaaliteltta: Twitter Politiikan Areenana Eduskuntavaaleissa 2015. Teoksessa Grönlund, K. & Wass H. (toim.) *Eduskuntavaalitutkimus 2015: Poliittisen Osallistumisen Eriytyminen,* 117–37. Helsinki: Oikeusministeriö.

Merry, M. K. (2014). Broadcast Versus Interaction: Environmental Groups' Use of Twitter. *Journal of Information Technology & Politics* 11(3): 329–44.

Nelimarkka, M., Laaksonen, S.-M., Marttila, M., Kekkonen, A., Tuokko, M. & Vilii, M. (julkaisematon käskikirjoitus). Influencing the agenda through social media: Online agenda building and normalization during a pre-electoral campaign period.

Stromer-Galley et al. (2016). Illuminating 2016. http://illuminating.ischool.syr.edu/

Sund, R. (2003). Utilisation of Administrative Registers Using Scientific Knowledge Discovery. *Intelligent Data Analysis* 7(6): 501–19.

Sund, R. (2015). Miksi isoon dataan hukutaan? *Tieto & Trendit – Talous ja hyvinvointikatsaus* 2/2015: 40-45.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. & Kappas, A. (2010). Sentiment in Short Strength Detection Informal Text. *Journal of the American Society for Information Science and Technology* 61(12): 2544–2558.

Wikipedia 2017. Big data. Luettu 30.1.2017. https://fi.wikipedia.org/wiki/Big_data

Zhang, A. X. & Counts, S. (2015). Modeling Ideology and Predicting Policy Change with Social Media. Teoksessa *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15,* 2603–12. New York, New York, USA: ACM Press.